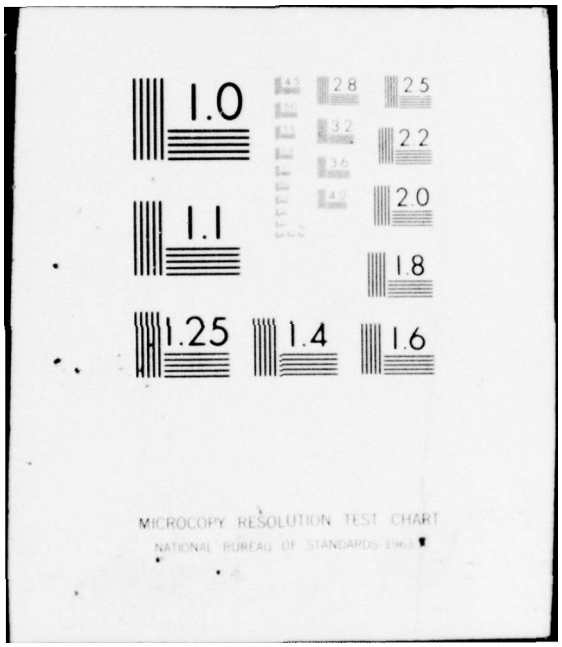
AD-A033 707 NAVAL POSTGRADUATE SCHOOL MONTEREY CALIF F/G 12/1
SASE VI AND THE STATISTICAL ANALYSIS OF SERIES IN EVENT IN COMP--ETC(U)
SEP 76 P A LEWIS
UNCLASSIFIED NPS-55LW76091 NL

1 OF 1
AD
A033707

END
DATE
FILMED
2-77


1.0
1.1
1.25
1.4
1.6
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963

NPS 55Lw76091

NAVAL POSTGRADUATE SCHOOL

Monterey, California

SASE VI AND THE STATISTICAL ANALYSIS OF SERIES

IN EVENTS IN COMPUTER SYSTEMS

by

P. A. W. Lewis

September 1976

Approved for public release: distribution unlimited.

Prepared for:

Chief of Naval Research
Arlington, VA 22217

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral Isham Linder
Superintendent

Jack T. Borsting
Provost

The work reported herein was supported by funds provided directly from the Chief of Naval Research under Grant NR-42-284.

Prepared by:

Peter Lewis

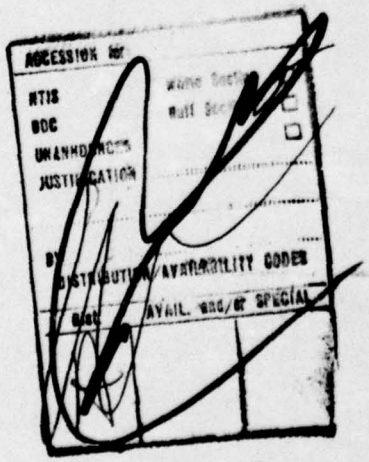
Peter A. W. Lewis, Professor
Department of Operations Research

Reviewd by:

Michael G. Sovereign, Chairman
Department of Operations Research

Released by:

Robert Fossum
Dean of Research



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: NPS55Lw76091	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Sase VI and the Statistical Analysis of Series in Event in Computer Systems,	5. TYPE OF REPORT & PERIOD COVERED: Technical Report,
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Peter A.W. Lewis	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Postgraduate School, Monterey, California 93940	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61153N, RR 014-05-01, Nooo1476WR60014
11. CONTROLLING OFFICE NAME AND ADDRESS: Officer of Naval Research, Arlington, Virginia 22217	12. REPORT DATE: Sep[illegible] 1976
	13. NUMBER OF PAGES: 25
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

RR01405 RR01405 01

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Statistical analysis	spectral analysis
Series of events	random variables
trends	point processes
renewal processes	cyclic trends
Poisson processes	computer systems

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We describe recent results in the development of methodology for the statistical analysis of univariate series of events (point processes), and some references to applications in the analysis and evaluation of computer system performance data. In addition, we describe the SASE VI program which has been developed to implement the methodology for the statistical analysis of series of events, in the monograph on this subject by Cox and Lewis. Various subroutines perform, among other things, tests for monotone and cyclic trends, tests for renewal and Poisson processes and two different

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601 |

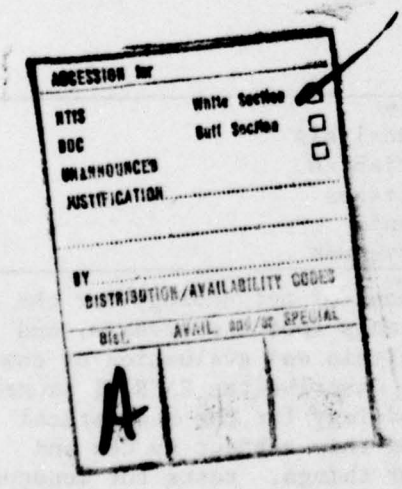
UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

251450

types of spectral analysis. The program can also be used to analyze any series of positive random variables such as counts of events in successive fired time intervals in a point process. It has been programmed in both FORTRAN and APL.

Multivariate series of events (point processes) present a much more difficult task and the methodology for their analysis has only recently been developed in a perforce fairly tentative manner. Applications in the analysis of computer system data and neurophysiological data are given. One problem here is the need for new data analytic methods for the analysis of data when trying to build models, and the lack of simple models for non-normal, positive multivariate time series. Some starts in these directions are described.



SASE VI AND THE STATISTICAL ANALYSIS OF SERIES IN EVENTS IN COMPUTER SYSTEMS

P. A. W. Lewis*

Naval Postgraduate School
Monterey, California

SUMMARY

We describe recent results in the development of methodology for the statistical analysis of univariate series of events (point processes) and give some references to applications in the analysis and evaluation of computer system performance data. In addition we describe the SASE VI program which has been developed to implement the methodology for the statistical analysis of series of events in the monograph on this subject by Cox and Lewis. Various subroutines perform, among other things, tests for monotone and cyclic trends, tests for renewal and Poisson processes and two different types of spectral analysis. The program can also be used to analyze any series of positive random variables such as counts of events in successive fired time intervals in a point process. It has been programmed both in FORTRAN and APL.

Multivariate series of events (point processes) present a much more difficult task and the methodology for their analysis has only recently been developed in a perforce fairly tentative manner. Applications in the analysis of computer system data and neurophysiological data are given. One problem here is the need for new data analytic methods for the analysis of data when trying to build models, and the lack of simple models for non-normal, positive multivariate time series. Some starts in these directions are described.

* Support from the Office of Naval Research under Grant NR-42-284 is gratefully acknowledged.

I. Introduction: The SASE IV Program

Series of events occurring randomly in time or space arise in many technological and scientific contexts. In statistics these event processes are called "stochastic point processes." The events may be the failures of a computer or the arrivals of nerve impulses at a synapse; the emissions of particles by a radioactive source or the errors occurring during transmission of binary data. An example is a study (Lewis, 1964) of computer failure patterns in time. Detailed statistical analysis (Lewis, 1964) assessed how one event type (maintenance) affected another (computer failure). It became evident even then that computer "reliability" would have to be broadened to include not only physical failures but also "congestion" failures.

Two examples of the analysis of series of events related to the problem of computer system performance evaluation are the analyses of a series of page exceptions by Lewis and Shedler (1973) and of transaction times in a data base system by Lewis and Shedler (1976). An analysis of page exceptions to two memory levels was given by Gaver, Lewis and Shedler (1974) and involved multivariate series of events; we discuss these later and concentrate first on univariate series of events.

SASE VI is an extension of the earlier SASE IV described by Lewis, Katcher and Weis (1969) and programmed in FORTRAN for IBM Model 360 and IBM Model 370 systems. SASE IV in turn was an outgrowth of SASE I and SASE II described by Lewis (1966). All the programs have been used on other manufacturers systems, particularly by neurophysiologists. SASE VI has extended plotting capabilities which make the data analytic examination of sets of data much simpler; otherwise it is similar to SASE IV. The extensions are discussed in the APPENDIX.

To understand the type of analysis performed by SASE VI, it is necessary to consider series of events in some detail. Such a series is shown schematically in Figure 1, events being represented by dots on the time axis.

In SASE VI, only univariate series of events are considered. The events can thus be distinguished only by the times at which they occur; other quantitative or qualitative data, such as the type of the event, are ignored for the purpose of analysis. Consequently the times-to-events T_i, with $0<T_1<T_2<T_3<\ldots$, or equivalently the times-between-events, X_i, can characterize the process.

There is, however, another equivalent way in which a series of events may be characterized. This is in terms of the counting process $N(t)$, the number of events occurring from the onset of measurement to time t. The representations of the series of events in terms of the times between events X_i and the number of events occurring in a given interval $N(t)$ are related by the following

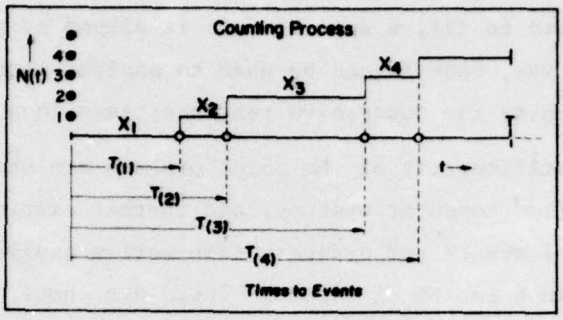


Fig. 1. Equivalent representations of series of events by the counting process $N(t)$ and the times-to-events $\{T_i\}$.

fundamental relationships, implicit in Figure 1:

$$N(t) < n \text{ if and only if } T_n = \sum_{j=1}^{n} X_j > t \tag{1}$$

and

$$\text{prob}\{N(t) < n\} = \text{prob}\{T_n > t\} \tag{2}$$

for $n = 1,2,\ldots$.

These relationships seem almost trivially obvious; yet they have important consequences. For example (if we are analyzing a series of events which are assumed to be stationary), a statistical analysis based on estimated correlation or spectral properties of the counting process, $\{N(t)\}$, is generally not equivalent to an analysis based on correlation or spectral properties of the interval process, $\{X_i\}$. Thus, if one is to explore a point stochastic process in detail, analysis should be made both of the number of events occurring over time and of the sequence of intervals separating one event from another. The exact interrelationship between the two types of analysis of stationary processes is implicit in (2) but is difficult to explicate; in certain cases it is clear which is most important. In superpositions, for instance, the spectral analysis of the interval process is much more informative than the spectral analysis of the counting process.

The SASE VI program, consisting of a main program and ten subroutines, carries out the computations for these two correlational analyses as well as some more particular types of analysis; for details, see Lewis, Katcher and Weis (1969) and the Appendix to this paper. The overall analysis differs from an ordinary time-series analysis in two chief respects: (i) the $\{X_i\}$ (interval) process is a time series in which the variables take on only positive values and do not have a Gaussian (normal) distribution, and (ii) the analysis of the $\{N(t)\}$ (counting) process has no direct counterpart in standard time-series

analysis. With regard to (i), a central role is played by the exponential distribution. Moreover, SASE IV can be used to analyze sequences of positive random variables such as the successive response times in a computer system.

Assumptions of stationarity of the point process are very iffy when analyzing data, for example, from computer systems, and further distinctions between the analysis of series of events and ordinary time-series analysis become evident in considering trends in a series of events. There are those which are best specified in terms of time, while others are most appropriately considered as functions of the event parameter i. Thus, computer failure-susceptibility might increase smoothly with time, affecting the intervals between events (failures) indirectly and in a complex fashion. On the other hand, if the events are the occasions when a rat presses a bar for food in a conditioning experiment, the increasing rate of events will be primarily a function of the number of previous bar pressings. Again time of day effects, e.g. effect of time of day on number of transactions in a data base system, are naturally specified through $N(t)$ and are not direct functions of serial number.

SASE VI includes a capability for detection of several types of trends. Thus, the TREND subroutine tests to determine if in a Poisson process the rate changes over time. (A Poisson process is an event process in which the number of events in non-overlapping time intervals are independent and have Poisson distributions.) Other subroutines (see Figure 2) specify further details concerning the nature and properties of the trend. Subroutine TREND also calculates quantities which can be used to test for trends using least-squares regression methods. These methods are flexible and allow various types of trends to be tested, working reasonably well under fairly weak assumptions as to the detailed structure of the series under study.

Where TREND does not indicate the presence of trends, stationary processes are involved. SASE subroutines (Figure 2) then test for fixed-rate Poisson processes; for correlation between intervals, i.e. for "renewal processes," in which the probabilistic structure of events and intervals "starts from scratch" after an event occurs. (A fixed-rate (homogeneous) Poisson process is one particular instance of a renewal process, but so are others in which the values of the variable are independent and have the same distributions.)

SASE VI subroutines also carry out spectral analysis for both the $\{N(t)\}$ and $\{X_i\}$ processes.

SASE VI has been widely used on a variety of computers with no apparent problems. This portability has been achieved at the cost of considerable sloppiness or looseness of programming.

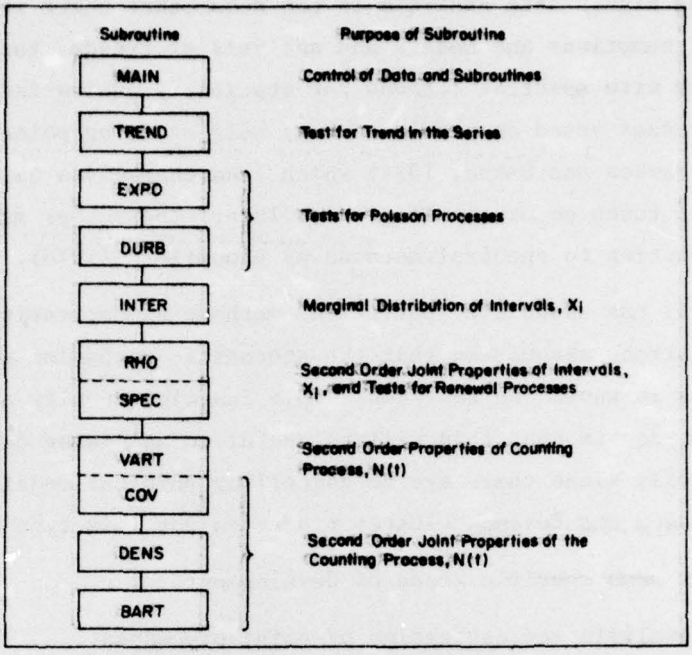


Fig. 2. SASE Subroutines

Several drawbacks in SASE VI include the need to load the whole program (approximately 250K bytes of core are needed), and the limitation to 2000 events. It is hoped to rectify these problems in a later version of the program. Also, several new developments in the statistical analysis of series of events which are described in the next section of this paper will be incorporated.

An APL version of parts of SASE VI has been written, but is not generally available. It offers several advantages over the FORTRAN version. I believe that with large workspaces and the input-output capabilities of recent versions of APL, this is the way to go. The interactive and array handling capabilities of APL make it ideal for data analysis.

II. Recent Developments

Recent developments in the statistical analysis of point processes have been summarized by Lewis (1972); see also Cox (1972) and Brown (1972). There is also a recent book on point processes by Snyder (1975) and a sequence of papers by Brillinger (1975a, 1975b). Brillinger's work is based heavily on spectral methods for stationary processes and has many points of contact with the work of Cox and Lewis. But Snyder (1975) does not reference any of Brillinger's work, and does not reference Cox and Lewis (1966). What then is the point of contact between these lines of development in the analysis of series of events?

The work of Cox and Lewis and the work of Brillinger are fairly complementary,

the former being highly data analytic in the sense that there is concern with validation of assumptions and models and analysis of trends, the latter being concerned mainly with spectral methods for stationary (univariate and multivariate) point processes based on models such as self-exciting point processes (Hawkes, 1972; Hawkes and Oakes, 1974) which lend themselves easily to spectral methods. I will touch on other differences later; the reader may be interested in a new introduction to spectral methods by Bloomfield (1976).

Snyder (1975) has based his statistical methods almost entirely on likelihood analysis and a strong assumption that the stochastic mechanism generating the series of events is known and that the sample function density can be written down. I have my doubts that this will be useful in analyzing data from computer systems, especially since there are no compelling physical models other than, in some cases, Poisson and Poisson cluster processes for this type of data.

Consider now some specific areas of development.

(1) Trend analysis and detrending of point processes.

In many fields of application, it has become increasingly apparent that stationary point processes are at best a convenient mathematical fiction. Most data exhibit fairly subtle trends and methods for testing for these trends are known (Cox and Lewis, 1966, Ch. 3); other data, however, exhibit gross trends, e.g. time-of-day effects in the series of arrivals at a queue, and techniques for the analysis and characterization of such data are only now beginning to be developed.

The situation is analogous to that in ordinary regression analysis and time-series analysis where one wants, for example, to estimate parameters in an assumed (linear) function for the mean, test the model for the mean function and then examine the model which is assumed for the residuals. The latter could include examining the residuals to test for independence, estimating the spectrum of the residuals and testing the assumed normality of the residuals. Techniques for these problems in the linear normal model are known (see e.g. Hannan, 1970).

By comparison, in point processes one might want to (i) estimate the rate function $\lambda(t)$, using either specific functional models or smoothing techniques; (ii) test specific functional models for $\lambda(t)$; (iii) detrend the point process, examine the 'residual' process and test the usual hypothesis that the events are generated by a homogeneous Poisson process.

When dealing with nonhomogeneous Poisson processes, the most appropriate detrending technique (Lewis, 1970, 1972) seems to be to transform the time-scale so that the $i^{\underline{th}}$ event occurring at time t_i now occurs at time

$$\tau_i = \int_0^{t_i} \hat{\lambda}(u)du,$$

where $\hat{\lambda}(u)$ is some estimate of the rate function $\lambda(t)$. Note that if $\lambda(t)$ is known, the τ_i's are a Poisson process. When $\lambda(t)$ is estimated from the data the distributional problems associated with determining properties of the $\{\tau_i\}$ processes are difficult.

Results for estimating parametric rate functions and given in Cox (1972) and Lewis (1972) and are developed in Lewis and Shedler (1976), where they are applied to the statistical analysis of transaction times in data base systems.

Fairly regular point processes are probably best dealt with by using log transforms of the intervals between events and then using ordinary time-series methods (Cox and Lewis, 1966, Ch. 3).

(2) Spectral analysis of point processes.

By spectral analysis of a point process (Bartlett, 1963) we mean the second-order spectrum of the counting function $N(t)$ (the count spectrum). Brillinger (1972) has put this spectral analysis on a firm footing in the context of a general spectral theory for stationary interval functions such as $N(t)$. He also proposed the use of higher-order spectra.

The spectral analysis may also be thought of as an ordinary second-order spectral analysis of a function $dN(t)$ which is a series of delta functions occurring at random times $\{T_i\}$ (see Lewis, 1970, for a heuristic interpretation). Note that it is not a second-order spectral analysis of the intervals between events $X_i = T_i - T_{i-1}$. The latter spectrum is useful for differentiating between renewal processes (for which it is flat) and non-renewal point processes. It may in fact be preferable to higher-order spectra of counts (Brillinger, 1972) in that it can be expected to exhibit fewer sampling fluctuations; in general my feeling is that a spectral analysis of the counts and the intervals should be tried before one goes to higher order spectra.

Note that a renewal process is completely specified by its second-order count spectrum. The spectrum is in fact essentially the Fourier transform of the renewal density or the intensity function.

One drawback to the spectral analysis of point processes is the large amount of time required for computation of spectral estimates. Only recently have French and Holden (1971), in an important paper, found a way to use the fast Fourier transform (FFT) in this problem. To illustrate their technique, consider the case when the series is observed for a fixed time t_0 and n events occur at times $T_1<T_2<\ldots<T_n$, where $T_n < t_0$. Note that n is the observed value of

the random variable $N(t_0)$.

Empirical spectral analysis of the point process is usually based on the finite Fourier-Stieltjes transform of $N(t)$,

$$H_{t_0}(\omega) = (\pi t_0)^{-1/2} \int_{t=0}^{t_0} e^{it\omega} dN(t) \tag{3}$$

$$= (\pi t_0)^{-1/2} \sum_{j=1}^{n} \exp(iT_j\omega) \tag{4}$$

$$= (\pi t_0)^{-1/2} \{ \sum_{j=1}^{n} \cos(T_j\omega) + i \sum_{j=1}^{n} \sin(T_j\omega) \}, \tag{5}$$

and the periodogram

$$I_{t_0}(\omega) = |H_{t_0}(\omega)|^2 . \tag{6}$$

The sum in (4) is not in the form required for the fast Fourier transform FFT algorithm (see, e.g. Cooley, Lewis and Welch, 1970, or Bloomfield, 1976), so that for each frequency for which the spectrum is required the number of operations required is of the order n. The periodogram is usually computed at frequencies ω_p such that $t_0\omega_p = 2\pi p$, where $p = 1,\ldots,P$, since the estimates are "well-behaved" at these points (Cox and Lewis, 1966, Ch. 5). Since P is of the order of n, the total number of computations is of order n^2.

Note that a point process can be sampled completely through knowledge of the event times, and the spectrum is not band-limited. However, French and Holden (1971) considered the realistic case where an upper frequency of interest $\omega_p = W$ was given, and considered the continuous time process

$$Y(t) = \int_0^{t_0} \frac{\sin W(t-\tau)}{2\pi(t-\tau)} dN(t) \tag{7}$$

$$= \sum_{j=1}^{n} \frac{\sin W(t-T_j)}{2\pi(t-T_j)} . \tag{8}$$

The Stieltjes convolution of $N(t)$ with the $\sin(Wt)/(2\pi t)$ function is equivalent to multiplying the spectrum by a frequency window which is flat to frequency W and zero thereafter. By sampling $Y(t)$ at intervals of width $\Delta T = \pi/W$, no frequency information is lost (aliasing) in the spectrum of $Y(t)$, which is the same as that of the point process up to frequency ω. The FFT can be applied to transform the sampled values of $Y(t)$; the main point in the technique, however, is that the sine functions have the same value for all t of the

form $\kappa\Delta t$, where κ is an integer. Thus, computation of the $Y(\kappa\Delta t)$'s is relatively simple and a considerable gain in computing speed shoudl be obtained.

There are some problems with this technique, e.g. it is not bias-free, as asserted by French and Holden (1971), but it appears to be of great value.

I will return to Brillingers higher-order count spectra when I discuss new models.

(3) Multivariate point processes

In almost all real cases we are interested in interactions between stochastic processes occurring at different places in space and time. Examples are the following:

(i) The spike trains (essentially point processes) occurring at the inputs and outputs of a neuron.

(ii) The successive response times in a data base system and the process of transaction initiation times at a certain point in the system.

(iii) Times of occurrence of page exceptions in a two-level memory multi-programmed computer system operating under demand paging (Gaver, Lewis, Shedler, 1972).

(iv) Times of occurrence of earthquakes at different points in California.

The examples are by no means exclusive but illustrate the main applications. We will only discuss multivariate point processes (Cox and Lewis, 1972) which might be thought of as a point process in which qualitative marks associated with each event can partition the set of events, although we note that Brillinger's work (Brillinger, 1972) encompasses more general situations. The analysis of multivariate point processes is discussed by Cox and Lewis (1966, Ch. 8; 1972), Perkel, Genstein and Moone (1966) and Brillinger (1972; 1975a, 1975b) and I will make only general comments here.

a. Dependencies between two stationary processes are usually handled via spectral methods, i.e. second-order cross-spectra which are then normalized to give quantities called coherences. This is Brillinger's approach, but it is not at all clear how useful second-order spectra are for point processes, which are of course a long way from normal processes in which the second-order spectra completely specify the dependence.

b. There is a problem of specifying what kind of dependency structures will occur in multivariate point processes. One can, for instance, generate many bivariate Poisson processes, defined to be bivariate

point processes in which the individual (marginal) processes are Poisson. A start at examining these structures has been given by Cox and Lewis (1972).

c. The above leads into questions of suitable models for multivariate point processes and this has been addressed in a tentative way by Cox and Lewis (1972) but is still wide open. Mutually exciting point processes (Hawkes, 1972) are being used but are very little understood and I believe are very untractable. Some of our recent work (Lawrance and Lewis, 1976; Jacobs and Lewis, 1976), when extended to multivariate processes, gives promise of much simpler models.

d. Spectral methods for examining dependencies in point processes are only one of many tools for doing this job and I have expressed my reservations above. There is a paper by Cox (1972) which is important and offers techniques based on likelihoods which needs further explanation. It might, in particular, be useful for analyzing dependencies in non-stationary point processes (see e.g. Lewis and Shedler, 1976); spectral methods are not applicable for non-stationary point processes.

e. Finally graphical, data analytic methods, rather than shotgun methods, should be used. This is particularly true since it is not completely evident that there is always stable stochastic structure in some of the processes encountered in examining computer systems.

An example of a very fruitful, simple graphical analysis of a bivariate point process occurred in Gaver, Lewis, and Shedler (1974). There a plot of the intervals between hits (page exceptions) to a lower-level memory were plotted against the number of hits to the higher level memory during that interval. The resulting plot is striking in that it consists (within statistical fluctuations) of two separate straight lines through the origin. The main thrust of the paper (Gaver, Lewis, Shedler, 1974) is to explain this structure.

This is really the way data analysis should be used; to suggest models or modify postulated models.

(4) Models

One aspect of the analysis of computer system data in the series of papers by Lewis and Shedler (1973); Gaver, Lewis and Shedler (1974) and Lewis and Shedler (1976) which comes home forcibly is that our techniques for analyzing time series data is heavily influenced by techniques for analyzing normal time series (c.f. Anderson, 1971). Thus one doesn't concern oneself much with the marginal distribution of the data, and second order correlation structure is all one looks at

unless one is looking at non-linear systems.

But the marginal distributions in most computer data are highly skewed positive random variables and are informative per se. Thus in the data on page exceptions in Shedler and Lewis (1973) a spectral analysis gave a clear indication of dependence between successive times between page exceptions and a quasi-oscillatory phenomenon which was then explainable. But the times between exceptions were very highly skewed and discrete. This gave rise to worry about the sampling theory of the spectral estimates, but a closer look at the empirical distribution function for the times between page exceptions showed that 5% of the times were exactly of length 1024. This was found to be a situation in which the program references sequentially each of the (4-byte) words in a (4k-byte) page and then goes on in a similar fashion to the next page.

Besides the fact such information from a data analysis is highly informative (and in fact more useful than a postulated model) analysis and modelling techniques for highly skewed, mixed and dependent random variables are non-existent. Lewis and Shedler (1973) used a 2-state univariate semi-Markov generated point process (Cox and Lewis, 1966, Ch. 7) to model the data, but estimation of parameters was perforce crude.

This has led us to derive new models which model the marginal distributions of the time series $\{X_i\}$ separately from its second-order correlation structure (Gaver and Lewis, 1976; Lawrance and Lewis, 1976; Jacobs and Lewis, 1976). We feel this is a promising start on a new methodology for analyzing data such as that arising in computer system evaluation studies. There are many open problems; as an example we have gone from a situation where there were no processes of dependent exponential random variables $\{X_i\}$ to one in which we have three such processes with exactly the same correlation structure of the type encountered in mixed moving-average autoregressive processes. Clearly higher-order spectra (Brillinger, 1972) will be needed to differentiate between the processes.

I may add that these new processes are easy to generate on computers and thus will lead to data-based models which will be highly useful in simulations of computer systems.

APPENDIX

THE SASE VI PROGRAM

The acronym SASE VI identifies the Fortran Computer Program for the Statistical Analysis of Series of Events, the sixth revision. SASE IV was distributed as a non-supported program through International Business Machines in 1969 and its procedures for usage are described by Lewis, Katcher and Weis (1969) which we refer to as LKW. The SASE IV program itself was correct and has proved to be a very useful tool, especially in the analysis of computer and neurophysiological data. The explanation for its use in LKW was relatively adequate. However, refinements in computation, listings and displays were deemed desirable as a result of usage on different data sets. In addition, changes have been made to facilitate its use simply as a program for the analysis of positive valued random variables. These occur if one does not record exact times of events, but counts of events in successive fixed time intervals.

The actual changes incorporated in SASE VI were programmed by Miss Patricia Combs of the Naval Postgraduate School staff, and the program is available from the author.

A. Description of the Program

The description of SASE IV remains valid except as noted here. The changes which are incorporated into SASE VI are listed in order of appearance in the program.

1. Main Program

Data may now have any input card format specified by the user on a parameter card described later. Restrictions on the data to be positive and to have no more than 1999 points still remain. The main program then tests all data and specified parameters for correctness and prints a listing of errors found. If errors are detected, the program execution terminates at this point. This particular feature may be annoying to the user whose low priority in the service queue gives him a long wait to find an error in a minor subroutine parameter, but it is felt that none the less this is desirable as rerunning of any subroutine involves a complete rerun of subroutine INTER; this is costly, especially with large data sets. Typical times to run the complete SASE IV program on an IBM Model 360/65 with 1999 points are 15 to 20 minutes. Other timing results are given in LKW.

2. Subroutine INTER

The subroutine is called automatically with each run of SASE. A new one page histogram plot and a smoothed estimated density function plot are automatically given in addition to the expanded histogram from SASE IV. This subroutine may, however, call any or all of three optional subroutines which assist in

analysis of the data. The first of the three is SECTION, a subroutine which divides the data set into the specified number of equally sized sections. The subroutine INTER then performs its complete analysis of each section and the statistical results of each section plus the statistical results for total set are tabulated for comparison. This option thus gives one an idea of trends or non-stationarities evident in the sample distributions from different segments of time.

The second option, subroutine TAIL, sections the complete set of ordered data (part of the output of subroutine INTER) into the specified number of tails, NT, then eliminates at each iteration, a minimum of $\lfloor N/NT \rfloor$ of the smallest values. The exact number eliminated varies because the subroutine will not cut a string of identical valued points; it searches the ordered data until a new unique value is located, then the remaining points are processed by subroutine INTER. The main value of this feature is the verification of an exponential tail (linear log survivor function) in the distribution of the data. This linearity, if present, gives some indication of a Poisson or compound Poisson model for the data.

The third option from subroutine INTER is JACKKNIFE. This feature computes jackknife statistics [Gray and Shucany, 1972; Miller, 1974] from chunks of the data from which successive equisized sections have been deleted, giving a total of NJ (a user specified number) sets of statistics. Jackknifing is particularly useful for small data sets to obtain variance assessments for estimated opoulation parameters.

3. Subroutine TREND

For each value of k, a plot of the values of the mean value and variance of data in the $i^{\underline{th}}$ set of k consecutive intervals has been added to subroutine TREND. This provides a visual as well as numerical listing of TREND indications.

4. Subroutine RHO

A new computation, its listing and its plot have been added to RHO. Specifically, it can be used to give the variance time curve for data which consists of counts of numbers of events occurring in fixed increments of time; otherwise it gives a variance-time sequence for the positive times between events in a series of events for which exact times of events have been recorded.

From Cox and Lewis (1966, pg. 72) we know that if the covariance between counts in intervals of size τ which have $J-1$, $J = 1,2,\ldots,$ identical intervals between them is $C_i(\tau)$, then the variance of counts in k contiguous intervals is

$$V(k\tau) = kV(\tau) + 2\sum_{i=1}^{k-1}(k-i)C_i(\tau) \qquad k = 1,2,\ldots . \tag{A.1}$$

From this we may use the already computed ρ_i, $i = 1,2,\ldots$ to obtain

$$C_i(\tau) = V(\tau)\rho_i$$

and hence $V(k\tau)$, $k = 1,2,\ldots$ For a sequence of independent variables $C_i(\tau)$ is identically zero for $i = 1,2,\ldots$ so that

$$V(k\tau) = kV(\tau) \qquad k = 1,2,\ldots . \tag{A.2}$$

i.e. the variance time curve is linear and passes through the origin. For counts this would be the case if we were dealing with a Poisson process. We notice that $V(\tau)$ is the variance of the counting random variable, $N(\tau)$, in one interval, $(t,t+\tau]$. The subroutine plots the sequence of values $V(k\tau)/V(\tau) - k$ for $k = 1,2,\ldots,M$ where M is the minimum of 200 or one half the sample size. For an independent sequence of random variables, i.e. $\rho_i = 0$, for all i,

$$V(k\tau)/V(\tau) - k = 2\sum_{i=1}^{k-1}(k-1)\rho_i$$

$$= 0 \qquad k = 1,2,\ldots .$$

Graphically, then, we are looking for deviations from zero in the plot. (The "variance time curve" values $V(k\tau)$ must be hand plotted using the printed values of $V(k)$.)

5. Subroutine SPEC

Two changes are incorporated into this subroutine which basically computes the spectrum of intervals. Of concern to the user only from the standpoint of efficiency, is a recursive form for computing sines and cosines. This change reduces SPEC running time by approximately 75%.

The second change is a convenience. The plots of the estimated spectrum are given separately for three values of smoothing. Previously, these were superimposed and sometimes barely distinguishable.

6. Subroutine DENS

The estimated intensity function, $\hat{m}_f(\tau)$, sometimes called the renewal function, is sensitive to the size of the sampling interval, SDT, specified by the user. In general, the optimal size is unknown in advance, so a loop has been incorporated into subroutine DENS which gives five separate plots of the estimated density using the following five interval sizes in estimation: SDT, 2×SDT, 3×SDT, 4×SDT and 5×SDT. This loop also places a new restriction on the size of SDT. Where SDT used to be

$$0 \le \text{SDT} \le t_{(n)}/L$$

now we must have

$$0 \le SDT \le t_{(n)}/5L.$$

Having five plots will save rerunning the program several times in an attempt to examine an apparent peak in the plot. Recall that $m_f(\tau)$ should be a constant value, m, for a Poisson process.

7. Subroutine BART

Subroutine BART has been recoded to allow computation of the (Bartlett) spectrum of counts by sections. Core requirements are thus reduced by 20,000 bytes for this subroutine. Additionally, a quadratic smoothing option has been added to the subroutine. The program user may select one of the following three options in BART:

(a) linear smoothing and plotting of all points;

(b) quadratic smoothing and plotting of all points;

(c) quadratic smoothing and plotting of only the midpoint value of an interval of size equal to the smoothing window.

A final convenience factor is that of automatic selection by default, i.e. non-specification, of most of the parameters needed by the subroutines. (The parameters listed are the same as those in the LKW manual for SASE IV.) Values chosen for this default feature are

$$K = 2$$
$$M1 = \max(2, [N/100])$$
$$M2 = 3\ M1$$
$$M3 = 5\ M1$$
$$DELT = \bar{x}/3$$
$$SDT = \bar{x}/2$$
$$L = \min\left\{\frac{N\ \bar{x}}{4\ SDT},\ 420\right\}$$
$$B = 2/(N-1)$$
$$P = 3N.$$

In addition, the value of the quadratic window in subroutine BART is chosen to be

$$KQUAD = \min\{100, [P/5]\},$$

if either of the quadratic smoothing options are selected.

Default options are provided only for convenience and the user should be careful to be sure that he knows what he is computing.

B. Use of SASE VI

The use of SASE VI remains essentially the same as the use of SASE IV. In order to avoid the necessity of two sources of information, we will list the complete control card sequence for calling sase VI and then some observations about various combinations of parameters. Details of parameters and computations are given in LKW.

The first control card is a 72 space Hollerith designation of the data set, e.g.

Feb 1972 hourly traffic count, location N3.

The second card lists, by means of a non-zero entry in the proper field, the subroutines to be called. They are

Column	Format	Subroutine
1-2	12	KDAT
3-4	12	TREND
5-6	12	RHO
7-8	12	SPEC
9-10	12	DURB
11-12	12	VART
13-14	12	COV
15-16	12	DENS
17-18	12	BART
19-20	12	IND
21-22	12	EXPO
23-24	12	# SECTIONS in INTER
25-26	12	# JACKKNIFES in INTER
27-28	12	# TAILS in INTER

In column 18 the entry should be 0, 1, 2 or 3 to obtain the desired selection from subroutine BART.

Entry	Selection
0	BART not selected
1	BART linear smoothing
2	BART with quadratic smoothing
3	BART with quadratic smoothing and only center points plotted.

Card three is used as follows:

Column	Format	Parameter	Range
1-10	I 10	N	$1 \leq n \leq 1999$
11-20	F 10.0	T	$T \geq 0$

Column	Format	Parameter	Range
21-30	I 10	K	$1 \le k \le N/20$
31-40	I 10	M1	$2 < M1 < 999$
41-50	I 10	M2	$2 < M2 < 999$
51-60	I 10	M3	$2 < M3 < 999$

The first two, N and T, must be specified explicitly in all cases; the others may be left blank; the program then selects the default values listed above.

Card four is used as follows:

Column	Format	Parameter	
1-10	F 10.0	DELT	$t_n/7999 \le DELT \le t_n$
11-20	F 10.0	SDT	$0 \le SDT \le t_n/5L$
21-30	I 10	L	$0 < L \le 420$
31-40	F 10.0	B	$0 \le B$
41-50	I 10	P	$0 \le P \le 6000$
51-60	I 10	KQUAD	$0 < KQUAD \le 200$

Card five is the input data card format statement. Starting in column 1 with a left parenthesis, the user lists the format and ends with a right parenthesis, for example,

```
(6(F10.4,10X))
↑
column one
```

C. Comments on Parameters

When the parameter k in subroutine TREND is small, the output from TREND will be quite lengthy especially when N is large, since it increases in N/k.

In subroutine BART, when KQUAD is small, the data window has strong negative side lobes which cause the smoothed estimates to be negative in some cases. It is recommended that KQUAD be greater than 50 if possible.

REFERENCES

Anderson, T. W. (1971). Statistical Analysis of Time Series, Wiley, New York.

Bloomfield, F. (1976). Fourier Analysis of Time Series: An Introduction, Wiley, New York.

Brillinger, D. R. (1972). "The spectral analysis of stationary interval functions," Proc. Sixth Berkeley Symp. Math. Statist. Prob. 1, Univ. Cal. Press: Berkeley, 483-513.

Brillinger, D. R. (1975a). "The identification of point process systems," Annals Prob. 3, 909-30.

Brillinger, D. R. (1975b). "Statistical inference for stationary point processes." In Stochastic Processes and Related Topics 1, M. L. Puri (ed.), Academic Press, New York, 55-99.

Brown, M. (1972). "Statistical analysis of non-homogeneous Poisson processes." In Stochastic Point Processes, P. A. W. Lewis (ed.), Wiley, New York, 67-89.

Cooley, J. W., Lewis, P. A. W. and Welch, P. D. (1970). "The application of the fast Fourier transform algorithm to the estimation of spectra and cross-spectra," J. Sound Vib. 12, 339-352.

Cox, D. R. (1972). "The statistical analysis of dependencies in point processes." In Stochastic Point Processes, P. A. W. Lewis (ed.), Wiley, New York, 55-66.

Cox, D. R. and Lewis, P. A. W. (1966). The Statistical Analysis of Series of Events, Methuen, London; Wiley, New York; Dunod, Paris.

Cox, D. R. and Lewis, P. A. W. (1972). "Multivariate point processes," Proc. Sixth Berkeley Symp. Math. Stat. Prob. III, Univ. Cal Press: Berkeley, 401-449.

French, A. S. and Holden, A. V. (1971). "Alias-free sampling of neuronal spike trains," Kybernetik, 8, 165-171.

Gaver, D. P. and Lewis, P. A. W. (1976). First order autoregressive Gamma sequences and point processes. To appear.

Gaver, D. P. Lewis, P. A. W. and Shedler, G. S. (1974). "Analysis of exception data in a staging hierarchy," IBM J. of Res. and Devel., Vol. 18, No. 5, 423-435.

Hannan, E. J. (1970). Multiple Time Series, Wiley, New York.

Hawkes, A. G. (1972). "Mutually exciting point processes." In Stochastic Point Processes, P. A. W. Lewis (ed.), Wiley, New York, 261-271.

Hawkes, A. G. and Oakes, D. (1974). "A cluster process representation of a self-exciting process," J. Appl. Prob. 11, 493-504.

Jacobs, P. A. and Lewis, P. A. W. (1976). "A mixed autoregressive moving-average exponential sequence and point process (EARMA 1,1)." Submitted to J. Applied Probability.

Lawrance, A. J. and Lewis, P. A. W. (1976). "An exponential moving average sequence and point process (EMA1)." To appear in J. Applied Probability.

Lewis, P. A. W. (1964). "A branching Poisson process model for the analysis of computer failure patterns," J. R. Statist. Soc. B, 1-59.

Lewis, P. A. W. (1966). "A computer program for the statistical analysis of series of events," IBM Syst. J. 5, 202-225.

Lewis, P. A. W. and Shedler, G. S. (1973). "Empirically derived micromodels for sequences of page exceptions," IBM J. Res. Devel. 17, 86-100.

Lewis, P. A. W. and Shedler, G. S. (1976). "Statistical analysis of non-stationary series of events in a data base system." To appear in IBM J. Res. and Development.

Perkel, D. H., Gerstein, G. L. and Moore, G. P. (1967). "Neuronal spike trains and stochastic point processes - II. Simultaneous spike trains," Biophysical J. 7, 419-440.

Snyder, D. L. (1976). Random Point Processes. Wiley, New York.

INITIAL DISTRIBUTION LIST

	Copies
Dean of Research Code 012 Naval Postgraduate School Monterey, CA 93940	2
Library, Code 0212 Naval Postgraduate School Monterey, CA 93940	2
Library, Code 55 Naval Postgraduate School Monterey, CA 93940	2
P. A. W. Lewis, Code 55Lw Naval Postgraduate School Monterey, CA 93940	10
Professors G. G. Brown, Code 55Bw	1
R. W. Butterworth, 55Bd	1
J. D. Esary, 55Ey	1
D. P. Gaver, 55Gv	1
K. T. Marshall, 55Mt	1
P. R. Milch, 55Mh	1
F. R. Richards, 55Rh Naval Postgraduate School Monterey, CA 93940	1
Statistics and Probability Program Office of Naval Research Attn: Dr. B. J. McDonald Arlington, VA 22217	3
Defense Documentation Center Cameron Station Alexandria, VA 22314	2
Technical Information Division Naval Research Laboratory Washington, D. C. 20390	6
Office of Naval Research New York Area Office 715 Broadway Attn: Dr. Robert Grafton New York, New York 10003	1
Director Office of Naval Research Branch Office 495 Summer Street Attn: Dr. A. L. Powell Boston, MA, 02210	1
Director Office of Naval Research Branch Office 1030 East Green Street Attn: Dr. A. R. Laufer Pasadena, CA 91101	1
Office of Naval Research Branch Office 1030 East Green Street Attn: Dr. Richard Lau Pasadena, CA 91101	1
Office of Naval Research San Francisco Area Office 760 Market Street San Francisco, CA 94102	1
Technical Library Naval Ordnance Station Indian Head, MD. 20640	1
Office of Naval Research Branch Office 1030 East Green Street Attn: Dr. D. Osteyee Pasadena, CA 91101	1
Naval Ship Engineering Center Philadelphia Division Technical Library Philadelphia, PA 19112	1
Bureau of Naval Personnel Department of the Navy Technical Library Washington, D. C. 20370	1
Director, Naval Research Laboratory Attn: Library, Code 2029 (ONRL) Washington, D. C. 20390	6
Director Office of Naval Research Branch Office 536 South Clark Street Attn: Dr. A. R. Dawe Chicago, IL 60605	1

	Copies
Library Naval Electronics Laboratory Center San Diego, CA 92152	1
Naval Undersea Center Technical Library San Diego, CA 92132	1
Applied Mathematics Laboratory Naval Ship Research and Development Center Attn: Mr. Gene H. Gleissner Washington, D. C. 20007	1